

The Power of R

In RStudio and it’s synergy with MS Power BI



22 May 2024

Just IT

Table of Contents

**Cleaning Data in RStudio2**

Step 1: Initial Exploratory Analysis2

Step 2: Clean Data4

Step 3: Further Exploratory Analysis6

Step 4: Exporting Data7

**Creating Power BI Dashboard with exported Data8**

Importing Data8

Dashboard5

**Reflective Learning 13**

**Cleaning Data**

**Step 1:**

Importing Movie DB csv into R. Using <- in R to associate the data table to “df”.

A screenshot of a computer

Description automatically generated

View() allows me to have a look at the data table “df” in another tab:

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

I can also do this by clicking on “df” in the environment.

Installed tidyverse package already so just needed to load this up. I can also activate tidyverse in the packages tab and check the box:

A screenshot of a computer

Description automatically generated

Using str() to identify structure of the data frames:

A screenshot of a computer program

Description automatically generated

Next to each column I can see char, int and num. Char is string characters, int is integer digits and num is the decimal values.

**Step 2:**

Checking null values:

A screenshot of a computer program

Description automatically generated

I can see there is 1 null value on Rotten Tomatoes, 1 null value in Audience..Score and 3 null values in Profitability.

Removed any null values in the data table:

A screenshot of a computer

Description automatically generated

I have to be careful as this can potentially remove all string entries if each null value associated with a unique column or primary key.

Used colSum() to check if the null values still existed in the table:

A screenshot of a computer program

Description automatically generated

Audience..score, Profitability and Rotten Tomatoes have come back as 0.

**Step 3:**

Creating a summary using summary():

A screenshot of a computer program

Description automatically generated

Code ran to generate a scatter plot:

A screenshot of a computer

Description automatically generated

The scatter plot shows the relationship between Lead Studios and Rotten Tomatoes scores, which can be double checked using the statistical summary generated earlier.

Code ran to generate a bar chart:

A screenshot of a computer

Description automatically generated

Bar chart shows the number of Films produced in each Year ranging from 2007 to 2011.

**Step 4:**

Exporting the cleaned data table as a new CSV file from RStudio:

A screenshot of a computer

Description automatically generated

Here I named the file “clean\_df.csv” having the csv extension so Windows recognises it as a CSV.

**Creating the Dashboard**

**Importing Data:**

Importing the CSV into MS Power BI by using get data.

A screenshot of a computer

Description automatically generated

I didn’t need to transform when I got the data as I have already done so using RStudio. I did however create new columns later in Table view to make the visuals look better, such as multiplying the Profitability column by one million, removing decimals and converting it into currency data type, so it is reflected in the dashboard.

A screenshot of a computer

Description automatically generated

**Dashboard:**

Link to my Dashboard:

<https://app.powerbi.com/groups/me/reports/22f450b1-6304-4255-9943-06391613e404/ReportSectioncb09664b857c0c575854?experience=power-bi>

The dashboard has a variety of charts that compares the different columns in the non-relational database:

A screenshot of a computer

Description automatically generated

I included a card slicer to allow the users to select which Lead Studio they wanted information for:

A screenshot of a computer

Description automatically generated

I included two drill through pages to cover stats for Genre and Lead Studio:

Genre and Film info for Disney include both Animation and Comedy.

A screenshot of a computer

Description automatically generated

As you can see below when selecting Animation there are 3 films Disney have made. The table shows which films they are and how much profit was made. The Pie chart indicate how many Animation films where made by which Company. When not using the card slicer this pie chart will have multiple segments, but here as Disney was selected 100% of the Pie is owned by Disney.

A screenshot of a computer

Description automatically generated

When not selecting Disney from the card slicer:

A screenshot of a computer

Description automatically generated

Lead Studio / Year looks drill through only displays when Lead Studio is present in the chart.

A screenshot of a computer

Description automatically generated

Here I utilised line graphs to track both the World Wide Gross and Profitability over years 2007 until 2011 as provided by the csv file.

A screenshot of a computer

Description automatically generated

**Reflective Learning**

I found the whole experience interesting and learnt quite a bit from this assignment. I didn’t need to transform the data as there were no relationships created due to the csv file being a NoSQL database and most of the data cleaning was done in RStudio.

Using RStudio was a bit difficult and without preset codes I would have struggled a bit. But I also recognise that viewing the denormlised data array in RStudio and being able to apply statistical analysis such as seeing which fields have null values and statistical distributions of each column without using Power BI was handy, especially if I wanted a simpler overview of the data. Also noticed that “No Reservations” film on line 39 didn’t have a respective Lead Studio so including the null values, some data seemed to be missing from the CSV.

Power BI was straightforward as I have used it before. I found the themes to be a bit difficult to work with as the client wants Blue, Green and Brown only, where as the themes included other colours, especially when a pie chart with many segments was involved. I would like to learn about other cool ways to display data in Power BI rather than using the same charts. I was able to create a pseudo interface with the card sliders so the user can select stats based on each Lead Studio. I tried to use the Gauge tile but it would always show the bar at 50% so might not be suitable for the data. I also included drill through pages to help display more stats that might be useful for user, one being about the film and the other about the finance statistics.

Overall pleased with the dashboard and feel a bit more confident using RStudio and Power BI.